Faculty of Cognitive Sciences and Human Development

A DEEP LEARNING APPROACH TO MALWARE DETECTION IN ANDROID PLATFORM

Corrine Francis

Bachelor of Science with Honours
(Cognitive Science)
2018

UNIVERSITI MALAYSIA SARAWAK

Grade: A

Please tick one

Final Year Project Report ☒

Masters ☐

PhD ☐

DECLARATION OF ORIGINAL WORK

This declaration is made on the 05 day of JUNE year 2018.

Student's Declaration:

I, CORRINE FRANCIS , 51559, FACULTY OF COGNITIVE SCIENCES AND HUMAN DEVELOPMENT, hereby declare that the work entitled, A DEEP LEARNING APPROACH TO MALWARE DETECTION IN ANDROID PLATFORM is my original work. I have not copied from any other students' work or from any other sources with the exception where due reference or acknowledgement is made explicitly in the text, nor has any part of the work been written for me by another person.

05 JUNE 2018

CORRINE FRANCIS (51559)

Supervisor's Declaration:

I, DR. ABDULRAZAK YAHYA AL-HABABI , hereby certify that the work entitled, A DEEP LEARNING APPROACH TO MALWARE DETECTION IN ANDROID PLATFORM was prepared by the aforementioned or above mentioned student, and was submitted to the "FACULTY" as a *partial/full fulfillment for the conferment of BACHELOR OF SCIENCE WITH HONOURS (COGNITIVE SCIENCE), and the aforementioned work, to the best of my knowledge, is the said student's work

Received for examination by: _____ Date: 05 JUNE 2018

(DR. ABDULRAZAK YAHYA)

I declare this Project/Thesis is classified as (Please tick (√)):

☐ **CONFIDENTIAL** (Contains confidential information under the Official Secret Act 1972)*
☐ **RESTRICTED** (Contains restricted information as specified by the organisation where research was done)*
☒ **OPEN ACCESS**

I declare this Project/Thesis is to be submitted to the Centre for Academic Information Services (CAIS) and uploaded into UNIMAS Institutional Repository (UNIMAS IR) (Please tick (√)):

☒ **YES**
☐ **NO**

Validation of Project/Thesis

I hereby duly affirmed with free consent and willingness declared that this said Project/Thesis shall be placed officially in the Centre for Academic Information Services with the abide interest and rights as follows:

- This Project/Thesis is the sole legal property of Universiti Malaysia Sarawak (UNIMAS).
- The Centre for Academic Information Services has the lawful right to make copies of the Project/Thesis for academic and research purposes only and not for other purposes.
- The Centre for Academic Information Services has the lawful right to digitize the content to be uploaded into Local Content Database.
- The Centre for Academic Information Services has the lawful right to make copies of the Project/Thesis if required for use by other parties for academic purposes or by other Higher Learning Institutes.
- No dispute or any claim shall arise from the student himself / herself neither a third party on this Project/Thesis once it becomes the sole property of UNIMAS.
- This Project/Thesis or any material, data and information related to it shall not be distributed, published or disclosed to any party by the student himself/herself without first obtaining approval from UNIMAS.

Student's signature: _____ Supervisor's signature: _____
Date: 05 JUNE 2018 Date: 05 JUNE 2018

Current Address:
FACULTY OF COGNITIVE SCIENCES & HUMAN DEVELOPMENT
UNIVERSITI MALAYSIA SARAWAK, 94300 KOTA SAMARAHAN, SARAWAK

Notes: * If the Project/Thesis is **CONFIDENTIAL** or **RESTRICTED**, please attach together as annexure a letter from the organisation with the date of restriction indicated, and the reasons for the confidentiality and restriction.

A DEEP LEARNING APPROACH TO MALWARE DETECTION IN ANDROID PLATFORM

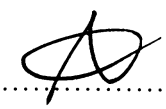
CORRINE FRANCIS
(51559)

This project is submitted
in partial fulfilment of the requirement for a
Bachelor of Science with Honours
(Cognitive Science)

Faculty of Cognitive Sciences and Human Development
UNIVERSITI MALAYSIA SARAWAK
(2018)

The project entitled 'A Deep Learning approach to malware detection in Android platform' was prepared by Corrine Francis and submitted to the Faculty of Cognitive Sciences and Human Development in partial of the requirements for a Bachelor of Science with Honours (Cognitive Sciences).

Received for examination by



.....

(ABDULRAZAK YAHYA SALEH AL-HABABI)

Date:

05th June 2018

GRADE
A

ACKNOWLEDGEMENT

First of all, I would like to thank God for giving me strength when I felt weak and wisdom when I lacked clarity so that I can complete this work. I thank Him for giving me everything that I needed the most.

The success and the final output of this project required a lot of guidance and assistance from many people and I extremely fortunate to have got this all along my journey to this project's completion. Whatever I have done is only due to such guidance and assistance and I would not forget to thank them. To my supervisor, Dr. Abdulrazak Yahya Saleh Al-Hababi, thank you for your time, effort and guidance that you have invested in me throughout this project. Thank you for believing me, encouraging me and you have nurtured me with your positive advices for all the time.

I would like to express my gratitude to my parents for kept on supporting me mentally and physically not just during the completion of this project but also during my whole studies in Universiti Malaysia Sarawak.

This project also cannot be completed without the effort and co-operation from the members under the supervision of Dr. Abdulrazak. In addition, grateful acknowledgement towards all my course mates who never give up in giving their support to me in all aspects of life and thank you for staying with me in ups and downs throughout all the year.

TABLE OF CONTENTS

LIST OF TABLES ..... **vii**
LIST OF FIGURES ..... **viii**
ABSTRACT ..... **x**
ABSTRAK ..... **xi**
CHAPTER 1: INTRODUCTION **1**
1.1 Overview 1
1.2 Problem Statement 5
1.3 Research Questions 6
1.4 Research Aims 7
1.5 Research Objectives 7
1.6 Research Scopes 8
CHAPTER 2: LITERATURE REVIEW **9**
2.1 Malware 9
2.1.1 History of Malware 11
2.1.2 The Purpose of Malware 17
2.1.3 Types of Malware 19
2.1.3.1 Trojans 21
2.1.3.2 Worms 22
2.1.3.3 Viruses 22
2.1.4 Ways of Spread 24
2.1.5 The Foundation of Anti-Virus (AV) Software 25
2.1.5.1 Top Three Anti-Virus Software Companies 27
2.1.6 Android Platform 31
2.1.6.1 Android Malware 33
2.1.7 The Analysis of gadgets Usage 34
2.1.8 The Attack of Android Malware 36
2.1.9 Methods of Prevention 37
2.2 Deep Learning 39
2.2.1 Famous Experts in Deep Learning (DL) Field 43
2.2.2 Types of Deep Learning 44
2.2.3 The Utilization of Deep Learning in Major Companies 44
2.2.4 The Relationship Between Deep Learning and Cognitive Sciences 47
2.2.5 The Relationship Between Computer Science and Cognitive Sciences 49
2.2.6 The Implementation of Deep Learning in Malware Prevention 50

2.3 Discussion 52
2.3.1 DL4MD (2015) 53
2.3.2 DroidDeep (2016) 53
2.3.3 DroidDelver (2016) 54
2.3.4 Deep Droid (2017) 55
2.3.5 Proposed Research 56
CHAPTER 3: METHODOLOGY 57
3.1 Introduction 57
3.2 General Framework 57
3.3 Data Pre-Processing 58
3.3.1 Data Preparation 58
3.4 Research Design 59
3.4.1 Extraction and Training 59
3.4.1.1 Long-Short Term Memory (LSTM) 59
3.4.2 Testing 69
3.4.3 Output 69
3.5 Evaluation 69
CHAPTER 4: FINDINGS 70
4.1 Introduction 70
4.2 Long-Short Term Memory (LSTM) Algorithm 71
4.2.1 Samples 71
4.2.2 Input Features 71
4.3 Codes 73
4.3.1 LSTM 73
4.3.2 Backpropagation 77
4.4 Result 83
4.4.1 Performance Analysis Based on Sample Sizes and Hyper-Parameter 84
4.4.2 Accuracy Detection Using Long-Short Term Memory (LSTM) Algorithm 86
4.4.3 Performance Comparison Between LSTM and Back-Propagation (BP) Algorithm 88
CHAPTER 5: DISCUSSIONS, RECOMMENDATIONS AND CONCLUSION 91
5.1 Introduction 91
5.2 Discussions 91
5.2.1 Research Aim & Objectives 92
5.2.2 Chapter's Explanation 93
5.3 Research Contributions 94
5.4 Future Work 94

5.5 Conclusion 95
REFERENCES **96**

LIST OF TABLES

TABLE 2.1: The timeline for malware's history (Adware, 2017; GData, 2017) 13
TABLE 2.2: Media influence in the malware's history (GData, 2017) 17
TABLE 2.3: The crucial events in 1987 26
TABLE 2.4: Methods of prevention 34
TABLE 2.5: Deep Learning influential experts 43
TABLE 2.6: Types of Deep Learning's descriptions 44
TABLE 2.7: Deep Droid (2017) Results and its comparison 55
TABLE 4.1: Error back-propagation code and its description 82
TABLE 4.2: Different number of epochs with different number of batch size 85
TABLE 4.3: Best hyper-parameter configuration on the training set 86
TABLE 4.4: Hyper-parameter for BP algorithm 89

LIST OF FIGURES

FIGURE 1.1: The top 10 malware (CIS, 2017) 2
FIGURE 1.2: Deep Learning (Santos, n.d.) 2
FIGURE 1.3: Facebook auto-tagging (Vaas, 2016) 3
FIGURE 1.4: AlphaGo (Lawler, 2016) 3
FIGURE 1.5: LSTM networks (Sonderby, Nielsen & Winther, n.d.) 4
FIGURE 1.6: News on cyberattack (Petroff & Larson, 2017) 5
FIGURE 1.7: Research questions 7
FIGURE 1.8: Research objectives 8
FIGURE 2.1: Statistical report of the Verizon data breaches 19
FIGURE 2.2: Types of malware (TSNC, 2016) 20
FIGURE 2.3: Eugene Kaspersky (Hern, 2017; Wikipedia, n.d.) 27
FIGURE 2.4: Steven Thomas (Griffin, 2008; Wikipedia, n.d.) 28
FIGURE 2.5: Steve Chang (Brewster, 2017; Trend Micro, n.d.) 30
FIGURE 2.6: Facts about malware (Matusevich, 2011) 32
FIGURE 2.7: The infection of Godless malware worldwide (Paganini, 2016) 34
FIGURE 2.8: Result of the research by Muduli (2014) 35
FIGURE 2.9: The comparison from 2012-2017 for new malware (TNW, 2017) 37
FIGURE 2.10: Deep Learning model 39
FIGURE 2.11: Types of learning in deep learning 40
FIGURE 2.12: Deep Learning's training process (Parloff, 2016) 42
FIGURE 2.13: Andrew Ng (McNulty, 2014) 43
FIGURE 2.14: Fei-Fei Li (Carey, 2014) 43
FIGURE 2.15: Geoffrey Hinton (McNulty, 2014) 43
FIGURE 2.16: Yann LeCun (McNulty, 2014) 43
FIGURE 2.17: Yoshua Bengio (Lowrie, 2016) 43
FIGURE 2.18: Google's Tensor Processing Unit (LinkedIn, 2018) 45
FIGURE 2.19: Interdisciplinary in Cognitive Sciences 48
FIGURE 2.20: Interdisciplinary of Cognitive Sciences and its description (Quist, n.d.) 49
FIGURE 2.21: Importance of computer science to CS (Thagard, 2017) 50
FIGURE 2.22: Facial recognition (Ayonix Corporation, 2017) 51
FIGURE 3.1: General Framework 57
FIGURE 3.2: Specific framework for Research Design 58
FIGURE 3.3: Derbin dataset 59
FIGURE 3.4: Four neural network layers (Olan, 2015) 60
FIGURE 3.5: Recurrent connection on the hidden layer nodes 62

FIGURE 3.6: Input signals framework moving through the hidden layers 62
FIGURE 3.7: Memory cell or cell state 63
FIGURE 3.8: Sigmoid layer 63
FIGURE 3.9: LSTM block illustration (Gibson & Patterson, 2017) 64
FIGURE 3.10: Training process 1 66
FIGURE 3.11: Training process 2 66
FIGURE 3.12: Training process 3 67
FIGURE 3.13: Training process 4 68
FIGURE 4.1: Experiment guidelines 70
FIGURE 4.2: Features extracted from the Derbin dataset 72
FIGURE 4.3: Forward propagation sub-parts 78
FIGURE 4.4: Back propagate error sub-parts 81
FIGURE 4.5: Train network sub-parts 82
FIGURE 4.6: Accuracy comparison by the size of epochs using 64 batch size and 5,000 samples 85
FIGURE 4.7: Accuracy comparison by the size of epochs using 64 batch size and 50,000 samples 86
FIGURE 4.8: LSTM model loss and accuracy comparison 88
FIGURE 4.9: Accuracy comparison of the LSTM and BP algorithms 89
FIGURE 4.10: Results description 90
FIGURE 5.1: Chapter 1-5 brief explanation 93

ABSTRACT

Throughout the years, mobile devices such as tablets, smartphones and computers are extremely widespread because of the development of modern technology. By using these devices, users all over the globe can easily accessed a huge range of applications from both commercial and private use. Malware detection is an important aspect of software protection. As a matter of fact, the development of malware had begun soaring as more and more unknown malware were discovered. Malware is a common term used to describe malicious software that can induced security threats to any devices and also to the Internet network. In this study, a malware detection that is based on Deep Learning approach that utilize the Long-Short Term Memory Networks (LSTM) model is utilized. The chosen approach will learn and train itself by using the features that are needed for malware detection using a large data sets for evaluating the trained algorithm. The performance of the model is evaluated by comparing it with the Back-Propagation (BP) model. Results that was achieved by conducting the necessary experiments proved that the LSTM model is capable to detect malware with the error loss of 0.6 and achieved an accuracy of 93.60% compared to BP with an accuracy of 82.857%.

ABSTRAK

Sepanjang tahun, penggunaan peranti mudah alih seperti tablet, telefon pintar dan komputer sangat meluas kerana perkembangan teknologi moden. Dengan menggunakan peranti ini, pengguna di seluruh dunia dengan mudah boleh mengakses pelbagai aplikasi yang besar dari kedua-dua kegunaan komersial dan swasta. Pengesanan perisian hasad merupakan aspek penting dalam perlindungan perisian. Sebagai hakikatnya, perkembangan malware telah mula melonjak kerana terdapat banyak malware yang tidak diketahui. Malware adalah istilah umum yang digunakan untuk menggambarkan perisian berniat jahat yang boleh menyebabkan ancaman keselamatan untuk sebarang peranti dan juga ke rangkaian Internet. Dalam kajian ini, pengesanan malware yang berdasarkan pendekatan Deep Learning yang menggunakan model Rangkaian Memori Jangka Panjang (LSTM) digunakan. Pendekatan yang dipilih akan menggunakan ciri-ciri yang diperlukan untuk pengesanan malware dan juga menggunakan set data yang besar untuk menilai algoritma terlatih. Prestasi model dinilai dengan membandingkan dengan model Back-Propagation (BP). Keputusan yang dicapai dengan menjalankan eksperimen yang diperlukan membuktikan bahawa model LSTM mampu mengesan malware dengan kehilangan kesilapan 0.6 dan mencapai ketepatan 93.60% berbanding BP dengan ketepatan 82.857%.

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

In January 2017, the detection of malware infection was at 43% and since then there has been a soaring activity of approximately 8% points starting from March and it extended to the following month. Then in April 2017, there are approximately 56% of new and unknown malware infection being detected as reported by the MS-ISAC (CIS, 2017). Figure 1.1 shows the top 10 malware that was detected by the MS-ISAC. Android malware has speeding up its infection activity where Google has issued an enormous security upgrade in July 2016 aiming 108 unprotected elements in Android platform (Zorz, 2016; DI, 2017). The security report encloses dozens of safeguards repairs for vulnerabilities for the Android system (Osborne, 2016). Smartphones possessed a huge pile of classified information that includes financial security and personal details. That quantity of valuable information can be retrieved illegally from the smartphones had made it the main mark for cyber lawbreaker from all stripes (DI, 2017).

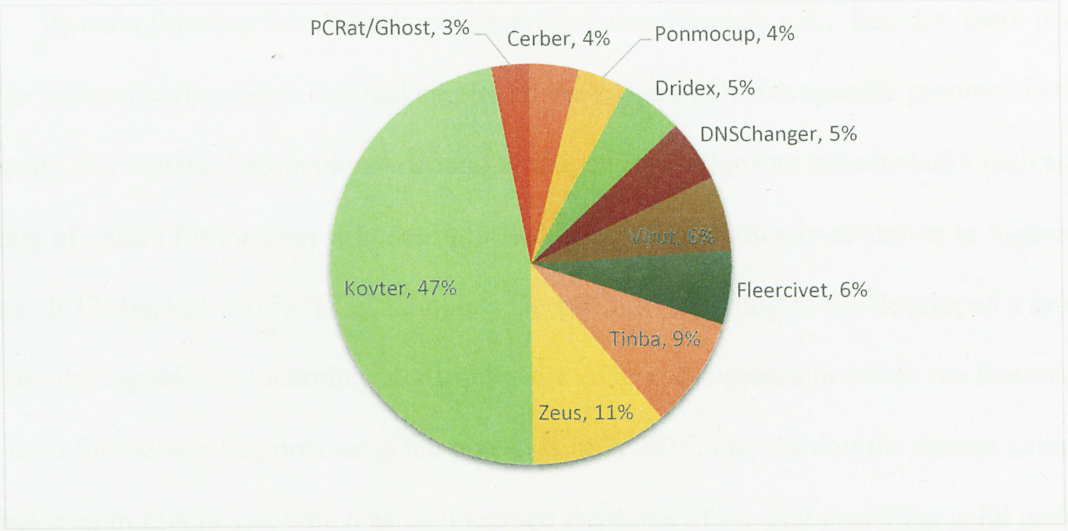


Figure 1.1 The top 10 malware detected in April 2017 (CIS, 2017)

Nowadays, there are numerous of research that corporate the utilization of Deep Learning (DL) technique in the malware detection for the Android platform. DL is a software that tried to imitate the activity in layers of neurons in the human brain, specifically in the neocortex segment that consists of 80% wrinkly part where the process of thinking happens (Hof, 2017). In addition, the other idea about DL also can be seen in Figure 1.2.

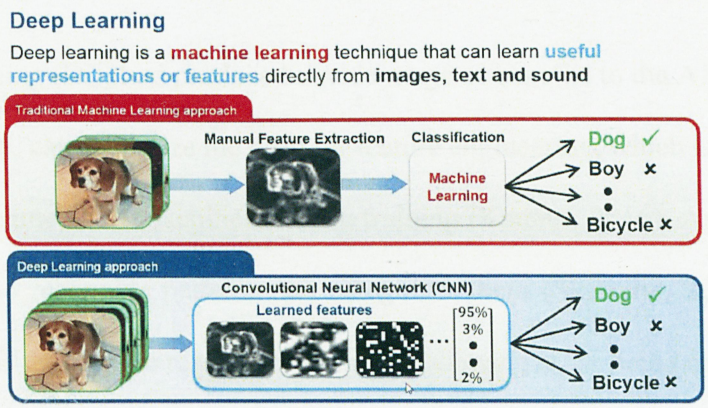


Figure 1.2 Deep Learning (Santos, n.d.)

Its main function is to learn how to represent data (Santos, n.d.). Besides, there is also a huge scale of corporations that had employed the usage of DL for specific products in their company. For instant, Facebook's Artificial Intelligence (AI) lab had execute tasks such as the tagging of names for the people in the uploaded images automatically as shown in Figure 1.3 (Metz, 2013; Becker, 2017). Then, Google's DeepMind Technologies has developed a system that has the capability of learning how to play the Atari video games in which the feature that was used for the learning process is the pixels. Within 2015, they exhibit the system called as AlphaGo as in Figure 1.4, which as well learned the game of Go that providing solid positive result where the system can defeat the expert Go player (Knight, 2016; Silver et. al, 2016).

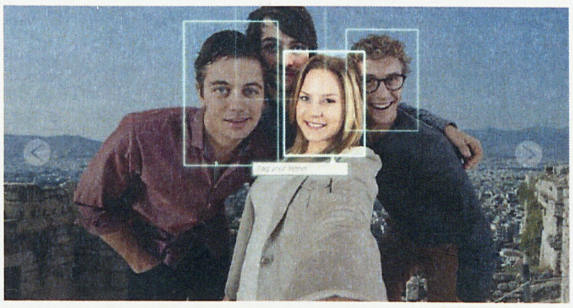


Figure 1.3 Facebook auto-tagging (Vaas, 2016)

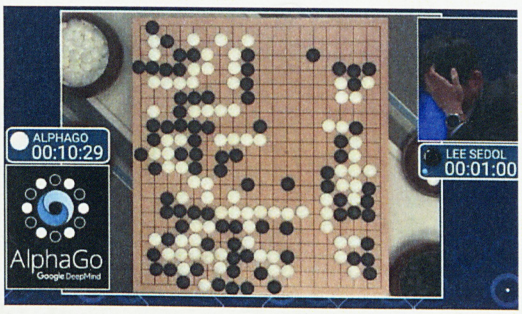


Figure 1.4 AlphaGo (Lawler, 2016)

Deep Learning (DL) had produced a wide range of benefits to the AI world and also to the human's life. DL can minimize the need for feature engineering, which is one of the widely time-consuming segments of machine learning training (Kadous, 2016). Apart from that, DL also can be used to recognize patterns and categorize them (Skymind, 2017) or in another context, the DL will inform the machine of what it's viewing at, or being cater for as a tide of numbers. In addition, DL can easily process unstructured data. For example, it does not need human interference to label everything in order to identify the sequence.

Long Short Term Memory networks (LSTMs) are a unique class of Recurrent Neural Networks (RNN) that has the capability of learning long-term interdependence (Olah, 2015) and also utilized the capabilities in arrangement forecast dilemma (Brownlee, 2017). RNN is another type of artificial neural network that is developed to identified designs in data arrangement, for example, handwriting, text and genomes (Skymind, 2017). LSTM technique is proposed by Hochreiter and Schmidhuber in 1997 and since then the technique has been polished and universalize in the associated fields (Olah, 2015). LSTM models have proven that it performs well than the RNN in language modelling (Sundermeyer, Schluter & Ney, 2012; Gers & Schmidhuber, 2001). Besides, there are numerous researchers that utilized LSTM in their research works (Azzouni & Pujolle, 2017; Sak, Senior & Beaufays, n.d.; Bakker, n.d.). Figure 1.5 shows the LSTM general frameworks.

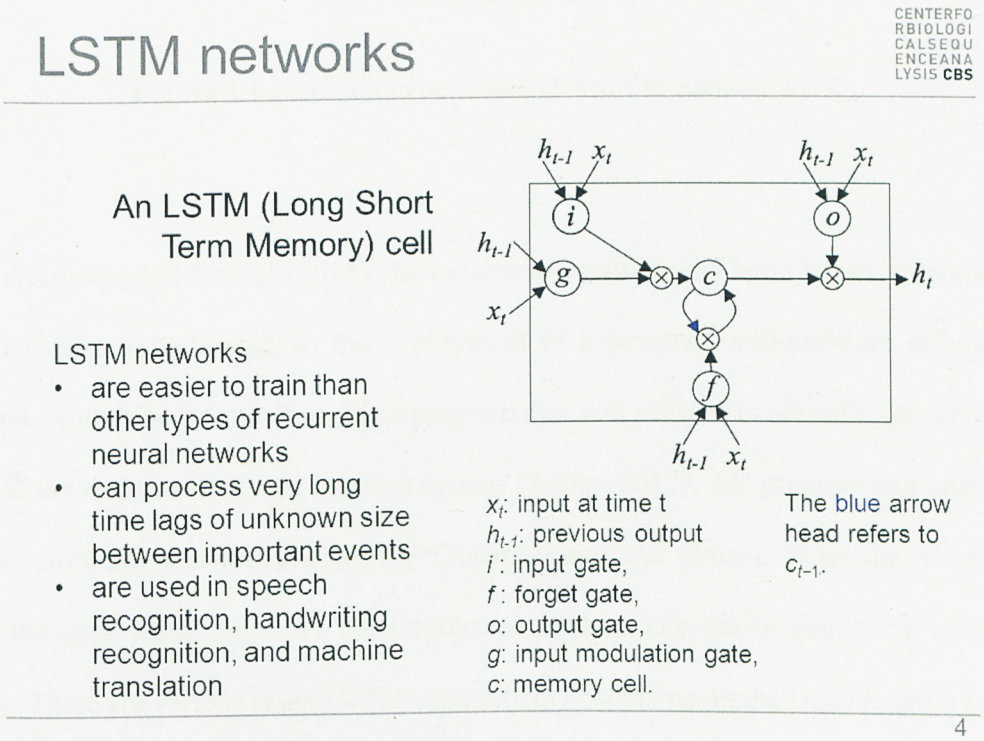


Figure 1.5 LSTM networks (Sonderby, Nielsen & Winther, n.d.)

1.2 PROBLEM STATEMENT

As the infections and attack cases of malware become quite serious throughout this year (Petroff & Larson, 2017; Newman, 2017) as shown in Figure 1.6, malware detection is crucial as it can help to prevent any unwanted problems that can bring catastrophic disaster towards the user life.

Hackers launched blistering attacks Tuesday, June 27, 2017 against companies and agencies across Europe. Major global firms are reporting they're under attack, including British advertising agency WPP, Russian oil and gas giant Rosneft and Danish shipping firm Maersk.

Figure 1.6 News on cyberattack (Petroff & Larson, 2017)

According to Metivier (2015), the number of malware had booming exponentially and becoming more complicated, so the deployment of a powerful anti-malware safeguards is important. Antivirus (AV) software is a program that was utilized to identify, quarantine, and get rid of the malware from the infected system (Soffar, 2017). AV program can protect user computer from any malware such as the Trojan, worms and viruses. Then, the AV will also protect the user when they surf the Internet as hackers can easily gain their information illegally. There are various research that researchers conduct using the Deep Learning method to detect malware in the Android platform (McLaughlin et al., 2017; Yuan, Lu & Zue, 2016; Yuan, Lu, Wang & Xue, 2014).

There is a necessary need for finding the solutions in order to resolve the problems that most of existing research experience frequently. The problems that needed to be focused more are the capability of achieving high accuracy detection and the low performances. Therefore, this study utilized the Deep Learning (DL) algorithm to solve the mention problems. The DL algorithm was widely used because it is an architecture that can fitting well to any fresh and unsolved dilemmas without difficulty (Kadous, 2016). In addition, the Long-Short Term Memory (LSTM) which is a type of DL will be utilized as the main algorithm in this study. The phrase of *long-short term* point to the reality that the LSTM is a model which can last for a long duration of time. According to Wikipedia (n.d.), this model is perfectly suitable for classification, processing, and prediction of time series. Thus, by applying the LSTM algorithm for malware detection in Android platform it can bring positive benefit such as providing a new opportunity for researchers to conduct research based on the LSTM algorithm.

1.3 RESEARCH QUESTIONS

To seek answers to these questions, this study is led by three research questions as shown in Figure 1.7. This is because a deep comprehension about the performance and the problems of the LSTM model in order to identify malware are necessary.

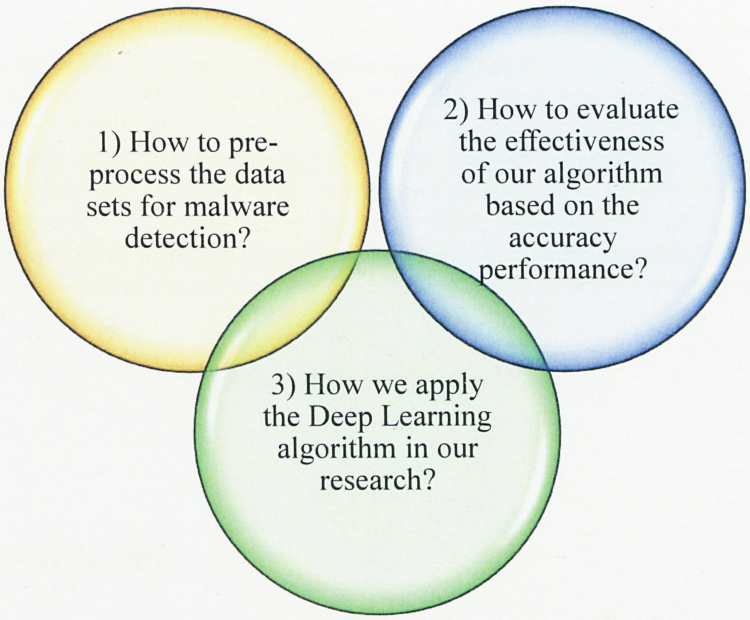


Figure 1.7 Research questions

1.4 RESEARCH AIM

The aim of this study is to successfully and efficiently identify malware in Android platform by achieving a high accuracy and also minimizing the error loss in the detection process.

1.5 RESEARCH OBJECTIVES

There are 3 main objectives in this study. All of the objectives are as shown in Figure 1.8.

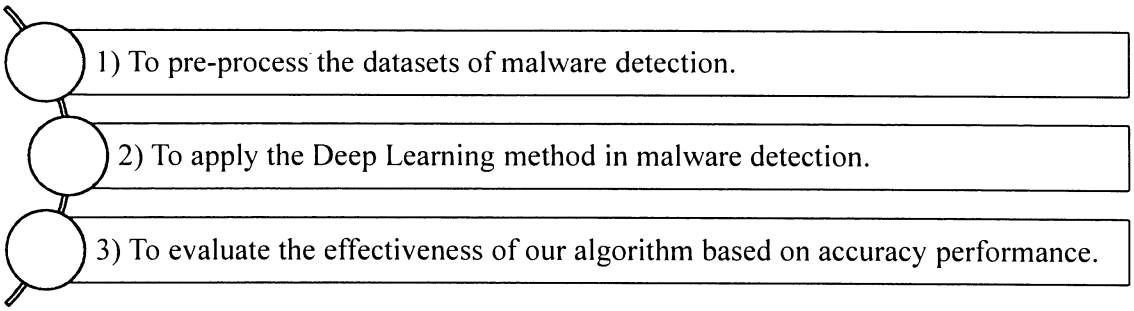


Figure 1.8 Research objectives

1.6 RESEARCH SCOPE

This study is conducted based on the research scopes as shown below:

a. **Dataset**: The Derbin dataset that is needed for this study were collected from the GitHub website (URL: https://github.com/prk54/malware-detection-machine-learning-approach)

b. **Deep Learning Technique**: The technique that utilizes by this study is the Long-Short Term Memory (LSTM).

c. **Software**: The software that is utilized in the study is the Spyder under the application of Python. The code that is use for the LSTM algorithm is based on the Python coding.

d. **Performance indicator**: In this study, the accuracy metric is utilized to evaluate the performance of the LSTM based on various experiments.

CHAPTER 2

LITERATURE REVIEW

2.1 MALWARE

Malware is a "malicious software" that were constructed to destroy or do unauthorized conducts on a computer system (TechTerms, n.d.). According to OECD (2007), malware is a common dub that being given for a prohibited component of software that being slid into a particular intelligence system without the system consent. Barraco (2016) stated that malware is primarily any computer program on a system that is not installed coincidently by the user or administrator. For a better understanding about the term of malware, The Network Support Company, TNSC (2016) has proposed in their site that malware is the short phrase for "malicious software" that composed of harmful programming that includes the code, scripts and etc. Many people had used the term virus to describe the malware when in fact, virus is actually one of the types of malware that can be found in any infected operating system. Norton (n.d.) stated that malware program is particularly formulated to acquire entry or ruin a computer without the possessor agreement or knowledge.

This malevolent software had very different types that only have one purpose that is to penetrate any computer without any concern at all. In addition, a software is commonly assumed as malware based on the true intention of the writer than its real hallmarks (Norton, n.d.). A program that basically act normal such as doing everything that they should be doing but did not tell the user what they was going to do is assumed as malware (Fisher, 2017).

According to Fisher (2017), the most common kinds of malware can look and act like a legal programs that the user use every day in their digital world. Malware can be identified by various indications. For example, Avast (2017) stated that a slow computer and repeated crashes is always considered that the user's computer is already infected with malware.

Nowadays, the "malware" statement have been a popular name utilized by computer experts to signify a various of types of unfriendly, invasive, or irritating computer system or program cryptograph. Malware is not the same or equal as flawed software but it is actually a software that turned a legal program that has a legal aims but consists of many malevolent bugs (programming blunders). The development of malware began to rise rapidly due to some aspects such as the temptation of money in which money can be made via systematic Internet crime (Norton, n.d.). According to Lee and Wright (2017), malware can performs various types of actions using a numerous of means into devices and networks. By using a USB drive, these harmful programs can be convey physically to any system that the USB being connected with. Then, a method that the malware always use as a medium of transportation to the internet is by drive-by downloads. Drive-by downloads is when the user automatically downloaded malevolent software into their computer's system without their consent (Lee & Wright, 2017).

There are some malware that can be remove via easy steps. These software can be removed by uninstalled it from the Control Panel, in the Windows operating systems (Fisher, 2017). Then, for some complicated malware that cannot be removed using the Control Panel, Avast (2017) recommended the user use anti-malware software. There are numerous of anti-malware that can be used to remove the malware inside a suspected computer. However, some malware that has its special approach of infecting computer needed a specific anti-malware that can remove the malware (Lemonnier, 2015). For the early prevention, Lemonnier (2015) stated that computer user can easily protect their computer by not clicking any doubtful emails, advertisement, links or website.